SCIENTIAMERICANA ,Revista Multidisciplinaria
Volumen 2 Número1, 2015

Medida de dispersión para variables nominales

Dispersion measure for nominal variables

Guillermo F. Parodi[1]

Artículo Recibido: 05/01/2015

Aceptado para Publicación: 15/02/2015

Resumen: Dado el carácter no numérico de las variables estadísticas nominales, tales como nacionalidad, sexo, etc., no es posible utilizar con ellas las medidas de dispersión habituales. Para obtener una medida apta para este tipo de variables, una forma es hacer uso del concepto de entropía de la teoría de la información y, sobre la base de ella construir una medida de dispersión.

Palabras Claves: Medida de dispersión, variables nominales

Abstract: Given the non-numerical character of the nominal statistical variables, such as nationality, sex, etc., it is not possible to use the usual dispersion measures with them. To obtain a measure suitable for this type of variables, one way is to make use of the concept of entropy of information theory and, based on it, to construct a measure of dispersion.

Keywords: Scatter measurement, nominal variables

Introducción

Dado el carácter no numérico de las variables estadísticas nominales, tales como nacionalidad, sexo, etc., no es posible utilizar con ellas las medidas de dispersión habituales.

Para obtener una medida apta para este tipo de variables, una forma es hacer uso del concepto de entropía de la teoría de la información y, sobre la base de ella construir una medida de dispersión.

En el parágrafo siguiente se revisarán los conceptos necesarios y posteriormente se definirá una medida de dispersión.

[1] Ingeniero Industrial, Docente investigador Universidad Americana.

Información y entropia

Dada una distribución de probabilidades o una distribución de frecuencias relativas, se define como información aportada por la observación del valor x_i, al valor dado por:

$$(1) \qquad I(x_i) = -\log_2(p(x_i))$$

La unidad de información se denomina bit.

Si se tiene $p(x_i) = 0{,}5$ el valor de información correspondiente será de 1 bit. Este resultado es coherente con la noción de bit en informática, en efecto al calcular la información que proporciona el conocer el estado de un dispositivo biestable, con probabilidades de estado iguales a 0,5, se obtiene el valor de 1 bit.

Cuanto menor es la probabilidad de un evento, mayor es la información que aporta al tener lugar. Un evento con probabilidad muy baja como por ejemplo $p=10^{-300.000}$ aporta una información de 10^6 bits.

Cuanto mayor es la probabilidad de un evento, menor será la información aportada por su ocurrencia. En el extremo un evento que ocurrirá con certeza (probabilidad =1), aporta una información nula, ya que su ocurrencia no implica ninguna nueva información.

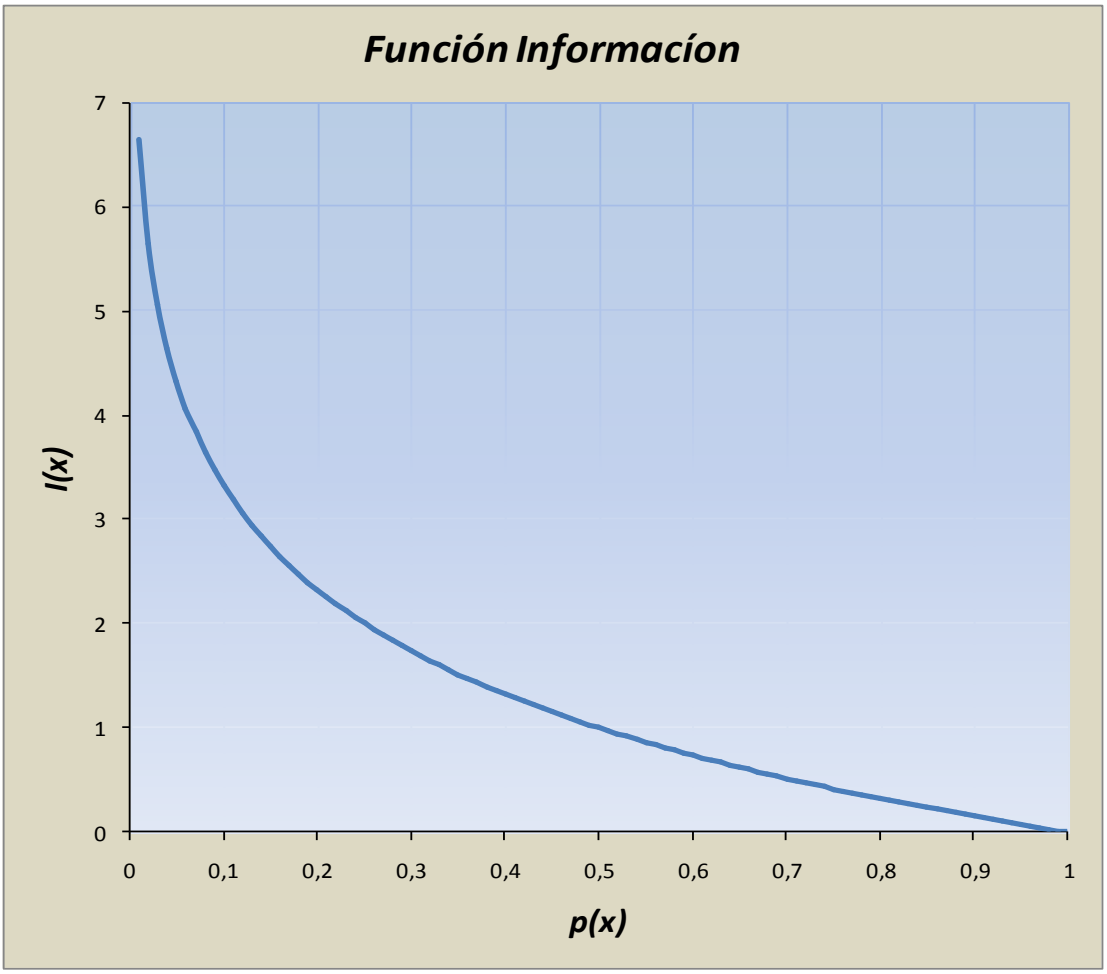


figura 1

La función entropía ***H(x)***, se define como el valor medio de la información extendida a todos los resultados posibles:

$$(2) \qquad H(x) = -\sum_{i=1}^{n} p(x_i).\log_2(p(x_i))$$

La función entropía tiene su máximo para todos los $p(x_i)$ iguales. Sabiendo que la suma de los $p(x_i)$ vale 1 y que deben ser todos iguales, se deduce que el valor máximo de H se

obtiene para todos los $p(x_i) = 1/n$. Reemplazando en (2), se obtiene:

$$(3) \qquad H(x)_{\max} = \log_2(n)$$

Medida de dispersión para variables nominales

Teniendo el valor máximo de H(x), ya es posible construir una medida de dispersión para variables nominales (DN), según la fórmula siguiente:

$$(4) \qquad DN = \frac{H(x)}{\log_2(n)}$$

Para todos los valores de $p(x_i)$ iguales, el valor de la DN vale 1 (dispersión máxima) y para una dispersión mínima, es decir un $p(x_i)=1$ y el resto cero el valor del DN es 0.

Observaciones prácticas

a) Normalmente las computadoras y las tablas permiten el cálculo de logaritmos neperianos (ln) o logaritmos de base 10 (log). Para el cálculo de logaritmos de base 2, pueden usarse las fórmulas siguientes:

$$(6) \qquad \log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$$

b)

$$(5) \qquad \log_2(x) = \frac{\ln(x)}{\ln(2)}$$

Cuando se tiene $p(x_i) =0$, el producto $p(x_i).\log_2(p(x_i)$ debe tomarse igual a 0 ya que:

$$(7) \qquad \lim_{x->0} (x . \log_2(x)) = 0$$

c) Para el caso de $p(x_i)$ = 1, el producto $p(x_i).\log_2(p(x_i)$ vale 0, ya que $\log_2(1) = 0$.

Demostraciones

Para encontrar el máximo de la función H(x) se hace necesario recurrir a la técnica de los multiplicadores de Lagrange.

Formando el Langrangiano:

$$(8) \qquad L = -\sum_{i=1}^{n} (p(x_i) . \log_2(x)) + \lambda.(\sum_{i=1}^{n} p(x_i) - 1)$$

Derivando a L respecto de $p(x_i)$ e igualando a cero, se obtiene:

$$(9) \qquad -\log_2(p(x_i)) + \lambda = \frac{1}{\ln(2)} \qquad i = 1,...,n$$

Derivando ahora respecto de lambda e igualando a cero, se obtiene:

$$(10) \qquad \sum_{i=1}^{n} p(x_i) = 1$$

La ecuación (9) por simetría implica $p(x_i) = cte$, que junto con la (10) da $p(x_i) = 1/n$. Reemplazando en la ecuación de H(x) (ecuación (2)), se obtiene $H(x)_{max}=\log_2(n)$.

Se trata de un máximo global pues la función entropía es la suma de funciones cóncavas y por lo tanto es cóncava.

En efecto:

$$y = -p(x_i).\log_2(x)$$

Por *(7) y=0* para *x=0*

Para x>0 la derivada respecto de *p(x)* será:

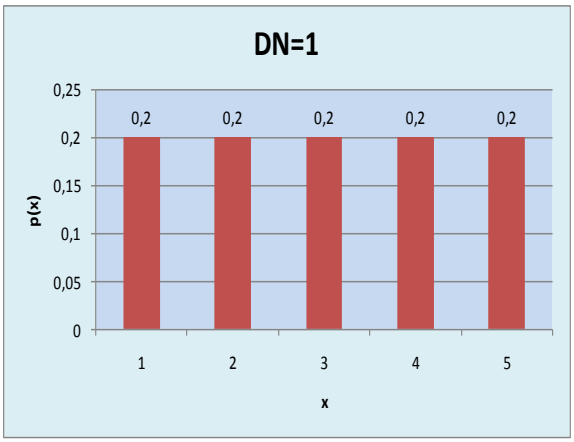
$$y' = -\log_2(x) - p(x).\frac{1}{p(x).\ln(2)} \qquad x > 0$$

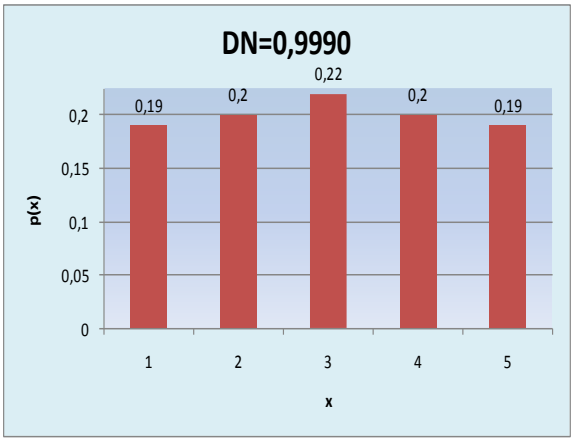
Y la derivada segunda:

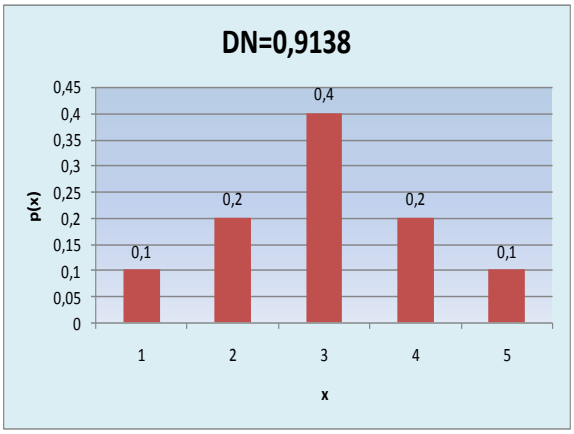
$$y'' = -\frac{1}{p(x).\ln(2)} = -\frac{1}{p(x)}.1{,}4426 < 0 \qquad x > 0$$

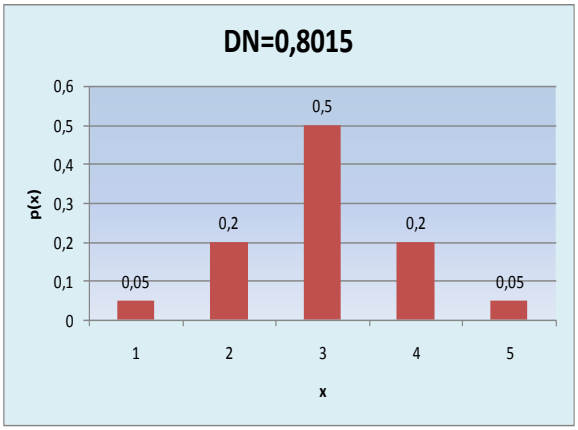
Por lo que y es cóncava y la suma de funciones cóncavas también es cóncava con lo que el punto estacionario hallado es un máximo global.

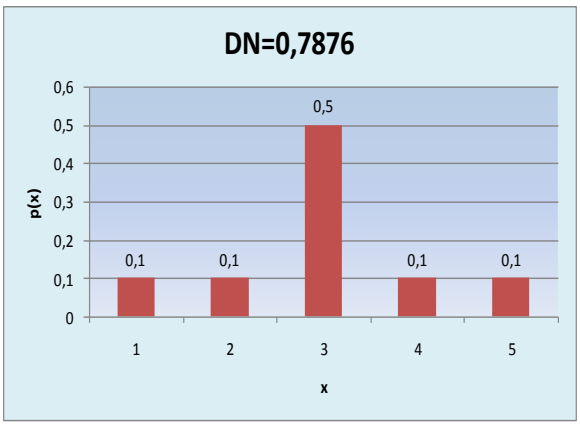
Ejemplos

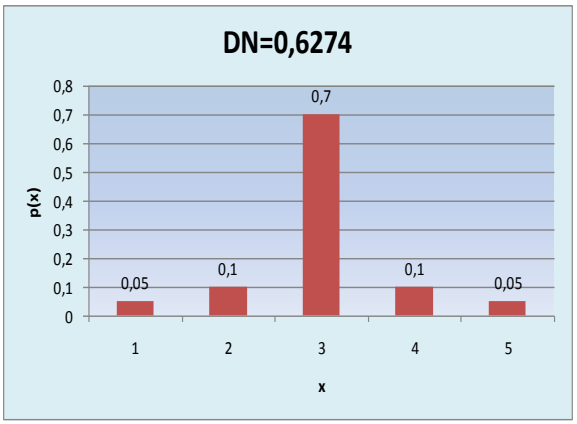

DN=1
0,25
0,2
0,15
0,1
0,05
0
p(x)
0,2
0,2
0,2
0,2
0,2
1
2
3
4
5
x

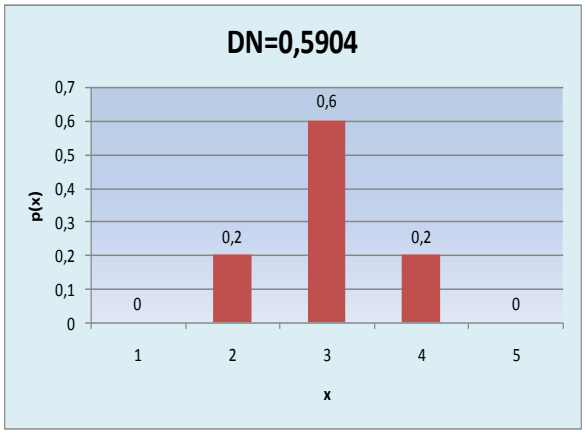

DN=0,9990
0,22
0,2
0,19
0,2
0,2
0,19
0,2
0,15
0,1
0,05
0
p(x)
1
2
3
4
5
x

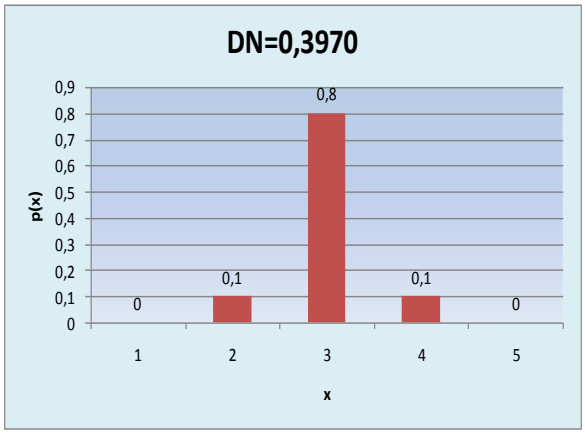

DN=0,9138
p(x)
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
0,1
0,2
0,4
0,2
0,1
1
2
3
4
5
x

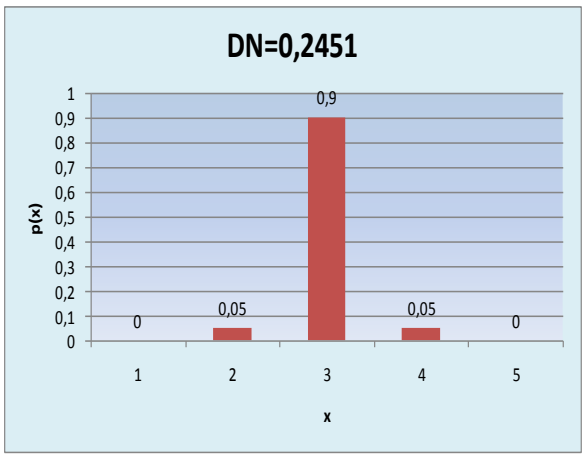

DN=0,8015
p(x)
0,6
0,5
0,4
0,3
0,2
0,1
0
0,05
0,2
0,5
0,2
0,05
1
2
3
4
5
x

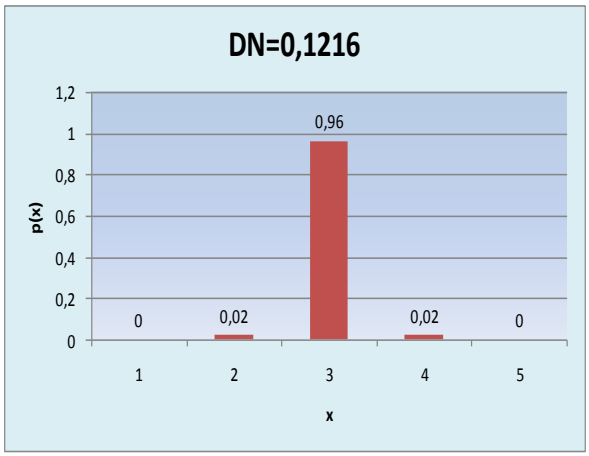

DN=0,7876
p(x)
0,6
0,5
0,4
0,3
0,2
0,1
0
0,1
0,1
0,5
0,1
0,1
1
2
3
4
5
x

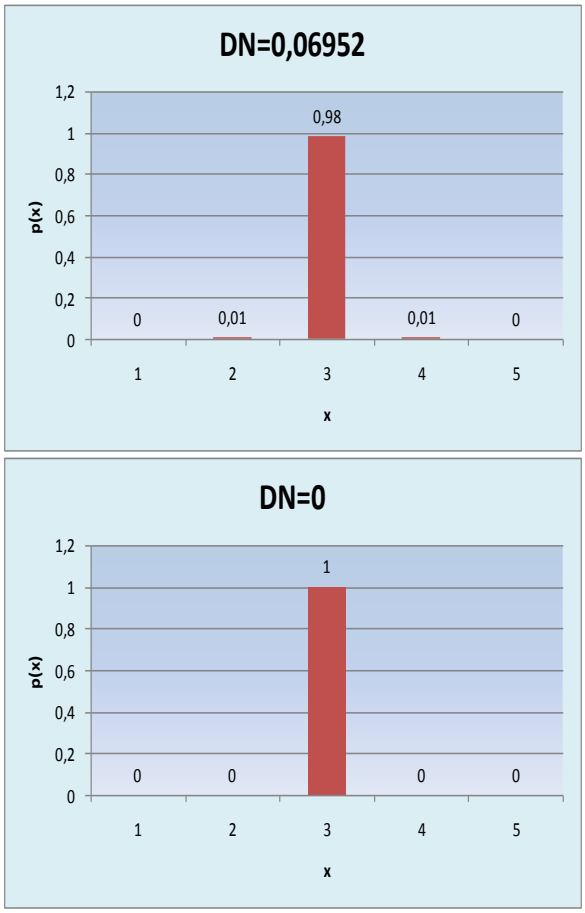

DN=0,6274
p(x)
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0,05
0,1
0,7
0,1
0,05
1
2
3
4
5
x


DN=0,5904
p(x)
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,2
0,6
0,2
0
1
2
3
4
5
x


DN=0,3970
p(x)
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,1
0,8
0,1
0
1
2
3
4
5
x


DN=0,2451
p(x)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,05
0,9
0,05
0
1
2
3
4
5
x


DN=0,1216
p(x)
1,2
1
0,8
0,6
0,4
0,2
0
0
0,02
0,96
0,02
0
1
2
3
4
5
x



NOTA: Pese a que se han dibujado gráficos simétricos, el valor de la DN es invariable con la permutación de los valores ya que la posición relativa no tiene influencia en el cálculo. Esta propiedad es deseada ya que el orden de las nominales es arbitrario.

Para las variables ordinales habría que buscar una medida que tenga en cuenta las posiciones relativas ya que en ese caso sí tienen importancia.

Referencia:

Elaboración del autor.